GMO INTERNET

InternetWeek2013
モバイル時代のインターネット～ソーシャルプラットフォーム設計最前線から～

# クラウドネットワーク編

GMOインターネット株式会社
中里　昌弘
masahiro-nakazato@gmo.jp

# アジェンダ

- 弊社ソーシャルアプリ用クラウドサービスの紹介
  - それを通じて思うこと
- ネットワークを構成する各パートで起きた事とその対処
  - トランジット
  - ピアリング
  - ルータ
  - ロードバランサ
  - L2ネットワーク
  - 購入
  - 機器選定
  - 連携
- まとめ

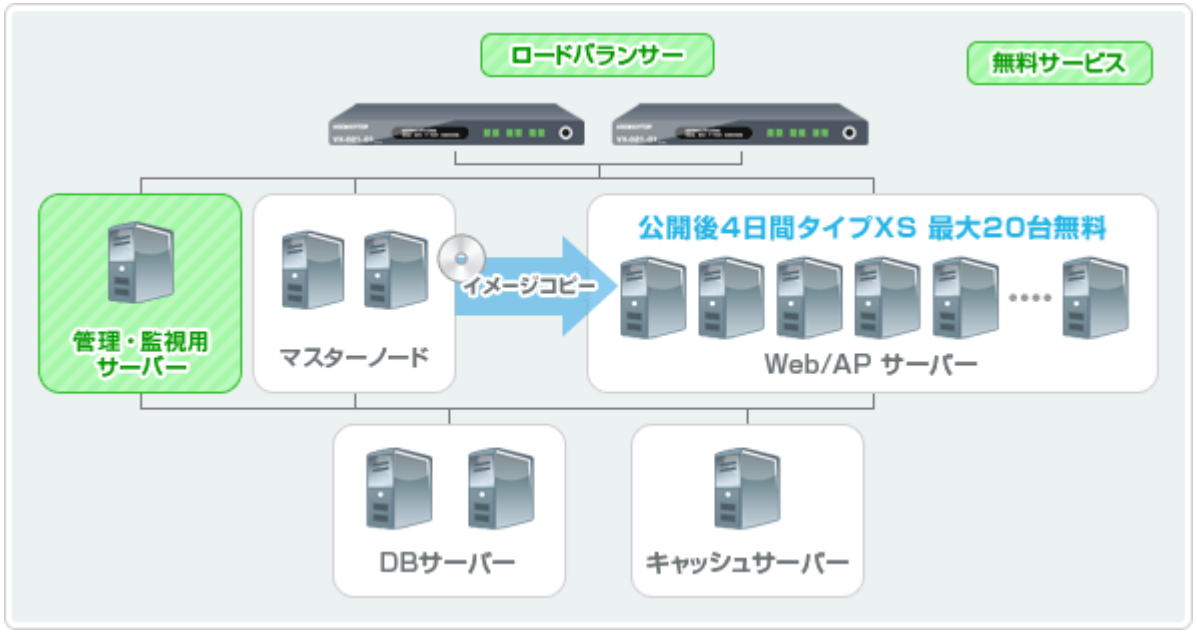
# はじめに - 弊社アプリクラウドの紹介

サービスマスコットの「美雲あんず」

## クラウドサービス名：GMOアプリクラウド

・GREE,Mobage,mixiなどのプラットフォームで公開する
　ソーシャルアプリの為のクラウドサービス

・アプリ開発・運用に不要な設備・機能を排除することで、コストパフォーマンスを実現



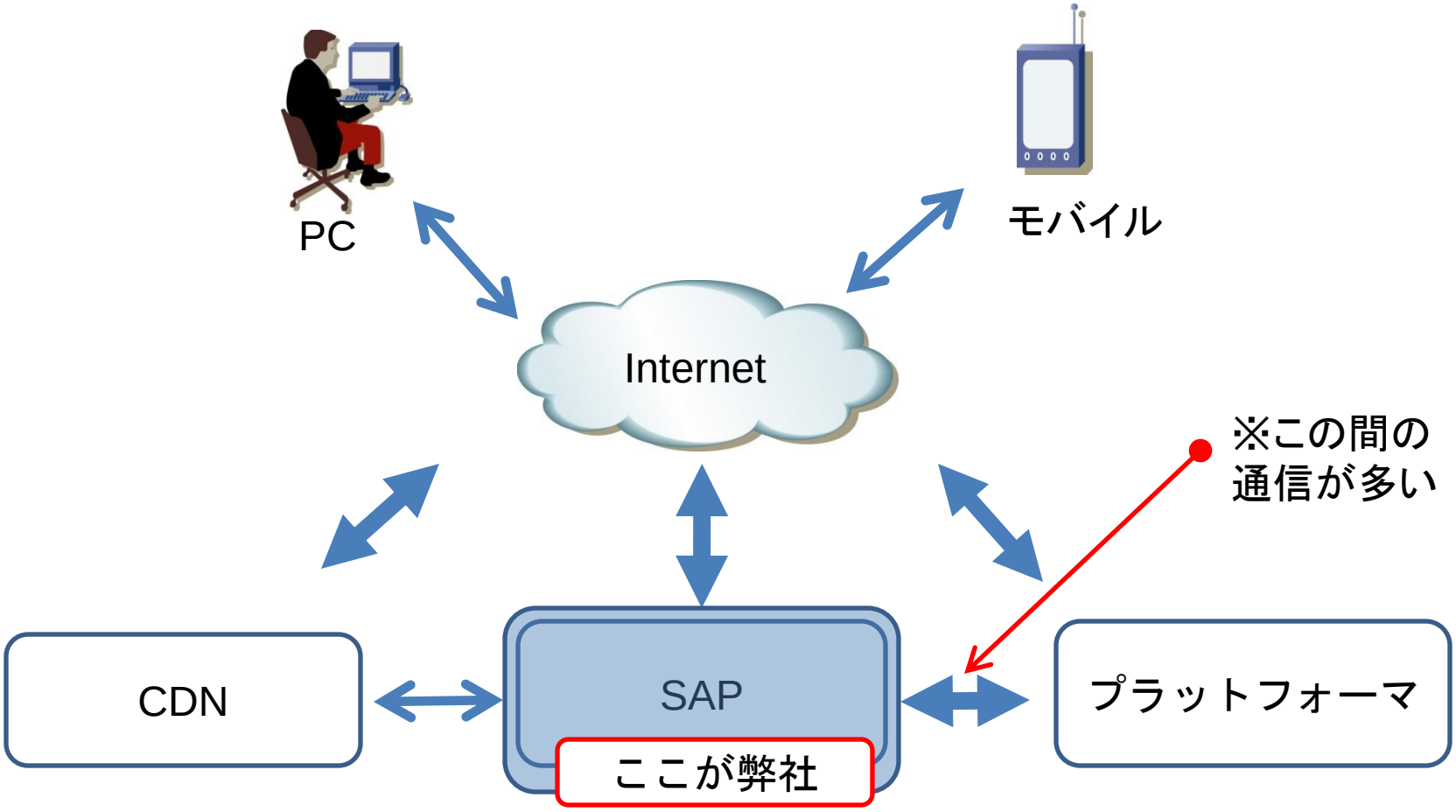
# SAP向けクラウドサービスを通じて思うこと

実はモバイルだから・・・というのはあまり意識せず
＝＞理由①へ

クオリティ、キャパシティ、コストパフォーマンスを
非常に意識するようになった
＝＞理由②へ

# 理由①:トラフィックの流れは固定ではない

SAPのゲームの作り方で通信経路は様々



# 理由②：事業者間の競争の激化

・売り物であるCPU,メモリ、ストレージについて
事業者間の比較が容易で、
使いたいときに使いたいだけご利用いただく
クラウドサービスの形態

＝＞スペック・コストが劣っていれば顧客獲得競争に
負けてしまう、またすぐに乗り換えられてしまう以前より厳しい世界

# 前提　売り物

クラウドサービスの売り物はcpuとメモリとストレージ(容量とIOPS)を紐付けたVM

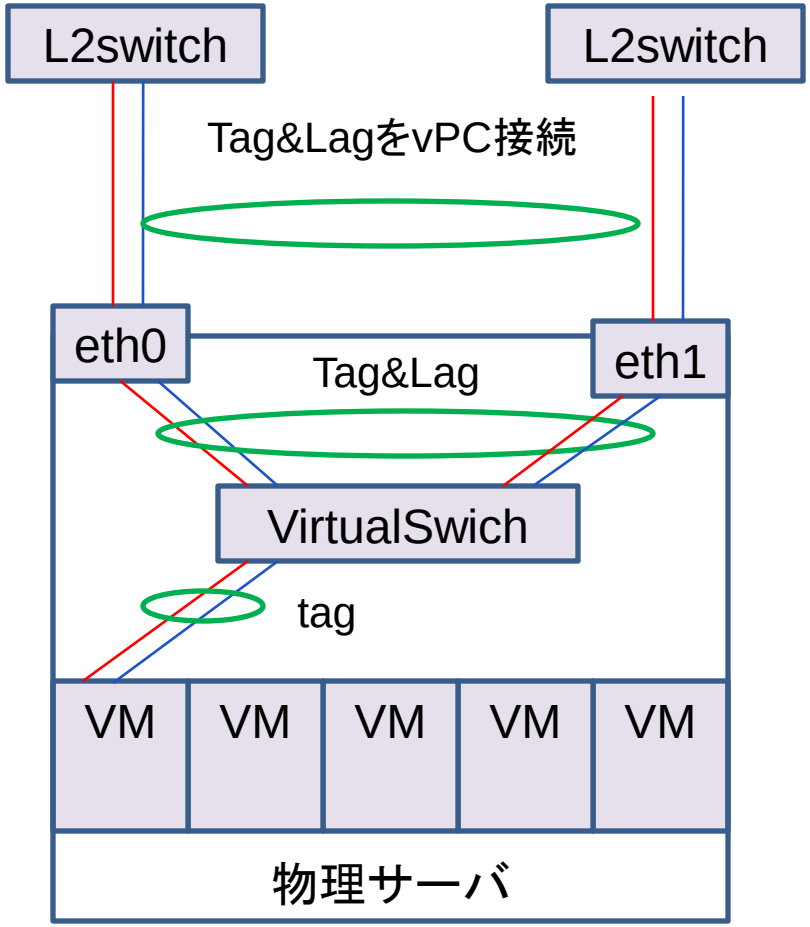
＝

| VM | VM | VM | VM | VM | VM |
|---|---|---|---|---|---|
| cpu | cpu | cpu | cpu | cpu | cpu |
| mem | mem | mem | mem | mem | mem |
| Storage | Storage | Storage | Storage | Storage | Storage |

・1Uサーバ1台あたり10～40VM程作れます

# 前提　サーバとの接続構成例

・1物理サーバあたり200-1000vlan程tagで通す




・2VLANで1顧客

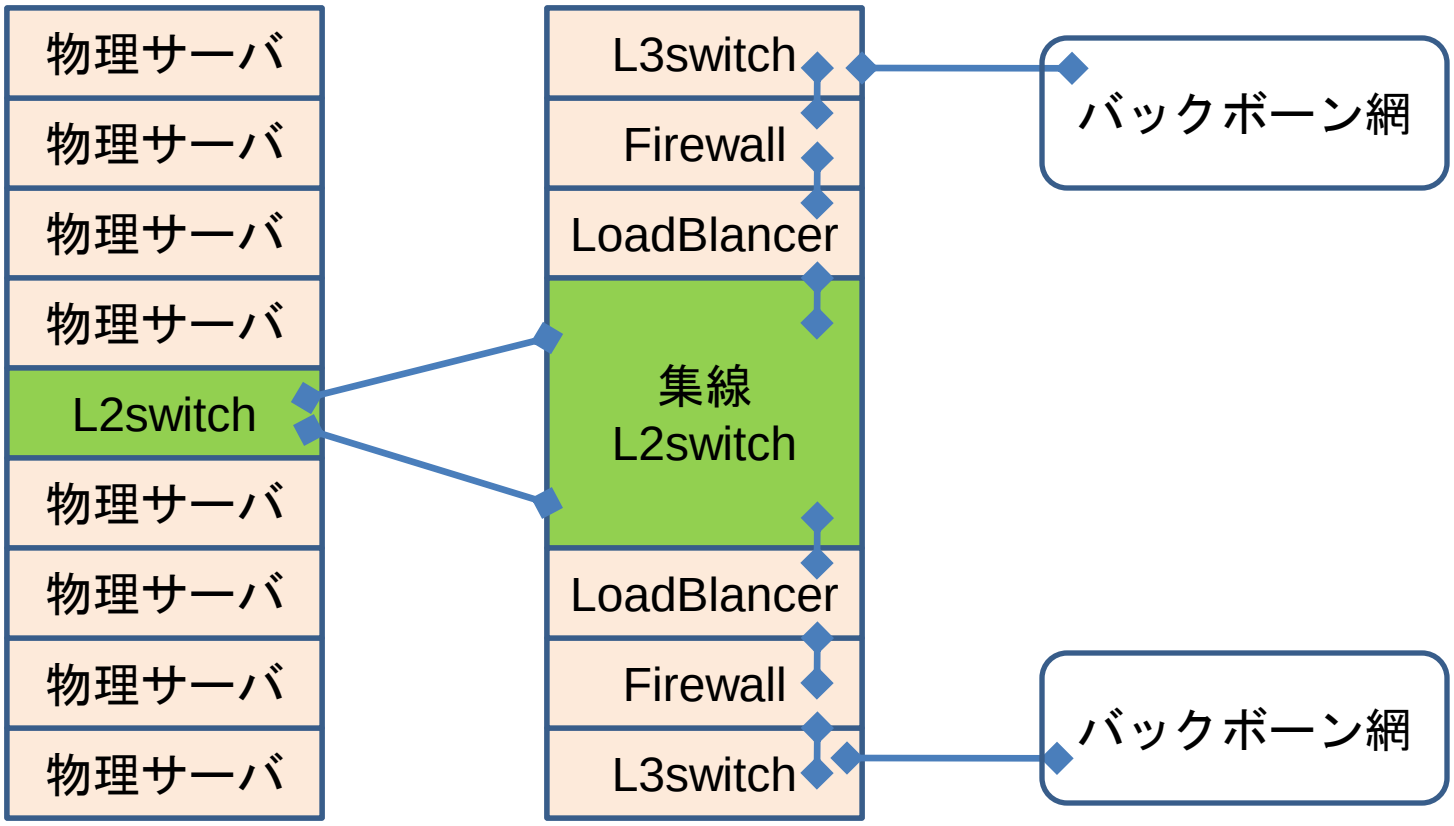
・物理サーバに多数のvlanを通すのはVMは多数ある物理サーバへランダムに生成される為

# 前提　1つのシステムのラック構成例

・サーバラック
20～100ラック程

・ネットワークラック
1-2ラック程

・1ラックあたり
サーバ25台



# クラウドサービスのネットワークって、、

上記の例で最大ものは、
100ラック×25サーバ×40vm=100,000vm
という売り物に対して、
1000vlanをtagで通すネットワーク構成をとる規模感
これを1つのユニットとする

複数のユニットをバックボーン網に接続している

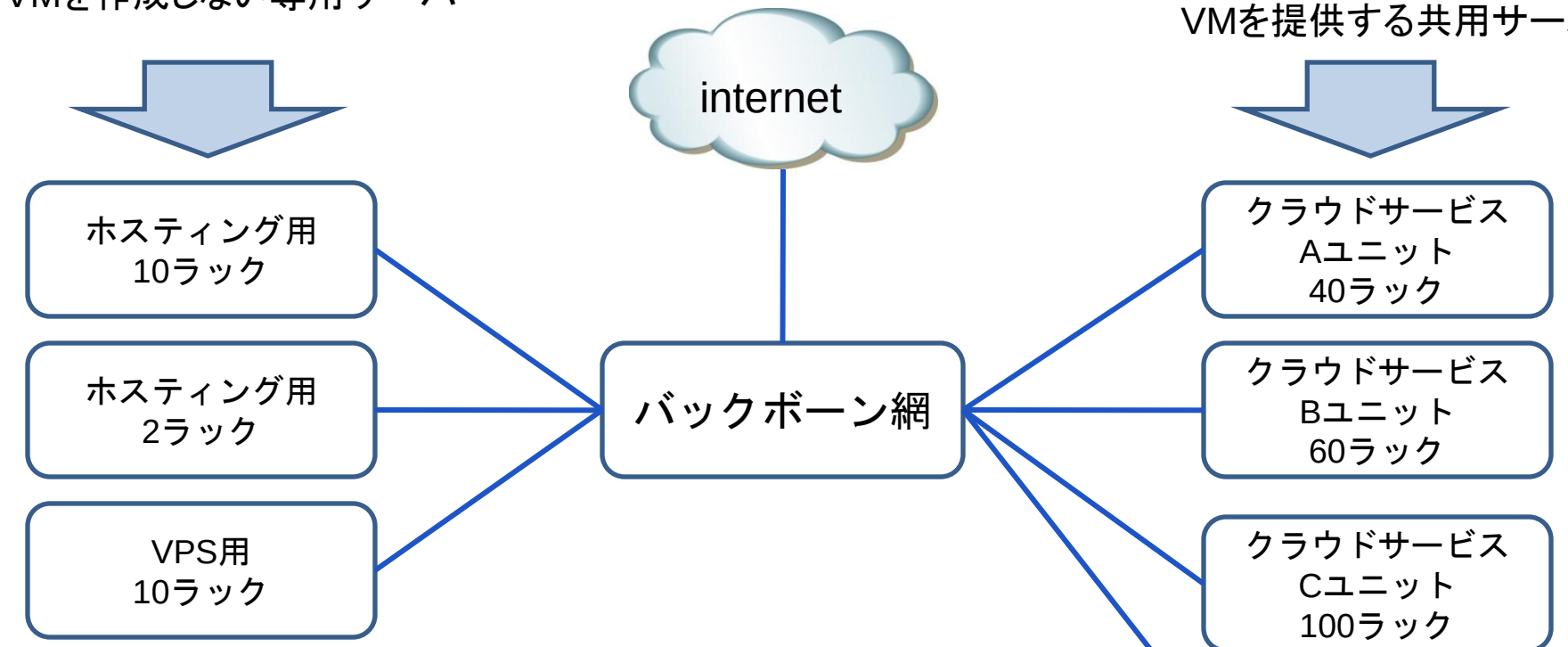
# 構成イメージとボリューム比較

## 従来サービス

ラック搭載数は低密度、
VMを作成しない専用サーバ

## クラウドサービス

ラック搭載数は高密度、
VMを提供する共用サーバ




マイグレーションして更にラックシェイプ傾向

新ユニットが続々と

# 次からは

過去に比較して

規模の大きなネットワークを抱えることになり

その中でネットワークを構成する各パートで起きたことを

ご紹介します

# トランジット >起きたこと

・遅延に敏感になる

=＞SAP<->プラットフォーマ間は通信断や遅延に非常にシビア

=＞モバイル網、プラットフォーマとの通信品質を高めることの商業的メリット

・トラフィックの予測が難しく、バーストが多い

=＞大型タイトルリリース等が原因

・クラウドサービス自体はDDoSは殆ど受けない

=＞アタックはホスティングサービス対象が殆ど

# トランジット >対処

・モバイル網を持つプロバイダを優先して選ぶ

=＞実際は選択の自由はあまり無かったり、乗換えが大変ですのでIX接続がある場合はそちらで代用する

・IX利用、プライベートピアを積極的に行い遅延やトランジット課金の要因を削る

=＞今は主要キャリアはピアを行ってくれる方向性

=＞トラフィックはピアでカバーする方向性　トランジットは保険と考える

・必ず毛色の違う2社以上と契約して通信の安定性確保とコスト競争を継続的にさせる

=＞選択の自由は武器になる

# ピアリング　>起きたこと

・IX経由のピアの信頼性

=＞トラフィックが流れすぎなのでピアを切れないか、なんていう打診があったことも

=＞そのトラフィックがトランジットに流れたら

・プライベートピアはルータのポートと専用回線調達などでコストが意外とかかる

=＞BGPルータのポート単価、データセンタ間の専用線はまだまだ高い

・ピア作業と調整は意外と工数を持っていかれる

=＞時間は有限！

# ピアリング >対処

・担当者の顔が見えるようにしてピアの信頼性を確保する

=>ピアは信頼関係を築き双方のメリットを出していきましょう

・担当者の顔が見えやすいIX業者を選択するのがベスト

=>そんな訳で某IX研究会に参加しています。

・プライベートピアよりもIXに回線を集中させたほうが回線とポート消費は少なく、管理も楽

=>このあたりは相手次第でケースバイケース

・IX業者のRouteServerは積極的に利用して工数を削減

# ルータ >起きたこと

・SAP<->プラットフォーマ間は通信断や遅延に非常にシビア

=>「この1分間中で何回か遅延がありましたね？」のような問い合わせも

・通信断を伴うメンテナンスが行いづらくなった

=>ospf等ルーティングプロトコルを使用した迂回もアウトな場合も

・BGPタイマ値は(30,90)でも遅い

=>「インターネットはベストエフォート」は、もはや通用せず

# ルータ　＞対処

・BGPはBFDを併用して断検知を3秒程にする

・BGP PICでコンバージェンスを短く

・OSPFはpoint-to-point等設定やarea内のノード数を減らすことでコンバージェンスを短く

・予算の許す限りいいルータを買う

# ロードバランサ　＞起きたこと

- バグが本当に多かった
  =＞稼働日数が一定になると通信停止、ospfがまともに動かない、mib取得でリブート…etc
  =＞あまりの多さにチームが破綻しそうになる

- 初期は受注毎に人が設定をしていた
  =＞ 作業量/ボリュームが多くチームが破綻しそうになる

- トラフィック量が凄まじい

# ロードバランサ　＞対処

- ベンダ、メーカとの長期バグフィックス体制の結果、
  安定バージョンで運用することが可能になる
  => ~~2年くらいかかりました~~

- 設定作業はAPI化してプロビジョニングシステムと連携させて
  人の作業を無くした
  => クラウドサービスの定常運用作業は人がしてはいけない！

- スイッチベースのパケット処理能力の高いロードバランサを
  並列構成を組むことによりトラフィックへの対処を行う
  =>　サーバベースのロードバランサは高機能だが使わない機能が多くコスト面、パケット処理能力でスイッチベースに軍配が上がる

# データセンター内ネットワーク >起きたこと

- 機器の上限値の数字に注意する必要あり
  =＞ mac address table 、lag数、logical interface(某C社固有のパラメータ)

- STPを利用するL2スイッチでは30ラック程が上限値
  =＞コンバージェンスに異様に時間か掛かるなどの不具合が出る

- STPトポロジの組み方も規模が大きいと注意が必要

- サーバと複数のスイッチを接続出来ないと単一障害点となる
  =＞ マルチチャーシラグ、vPCはほぼ必須

# データセンター内ネットワーク　＞対処

・最新(2013Q3)の~~b社c社~~fabric系switchを使用すると
　200ラック程の規模まで対応可能
　=＞動的vlan割り当て等のプロビシステムと連携をさせないとまずい場合も

・STPはループトポロジの輪を複数被せると逆に収束に
　時間が掛かる　フルメッシュにすればいいという訳ではない
　=＞ベンダでもノウハウを持っていないことが多く、
　　自分たちサーバを作成後の障害試験はで行うのがベスト
　=＞でももうSTPスイッチではクラウドサービス作らないよね・・？

・サーバ接続は1G UTPから10G　twinaxにシフト
　=＞従来スイッチは割高、メーカもお勧めしてこない
　=＞新しく構成するならfabric系switch

# データセンター内ネットワーク >対処

・ケーブル挿し間違えによるループ対策を必ず

=＞ bpduguardやedge loop detectionの設定はいれよう

=＞色分けやナンバリングを駆使してオペミス予防する

・機器のエアフローに注意　熱で故障率上がります

=＞全ての機器がデータセンタ向けではないので、機器選定時にエアフローを確認する
クセをつける

# 機器 –購入　＞対処

・メーカ、ベンダは必ず2社以上を徹底的に競わせる
　=＞ただし相手は人　win winの関係を築くよう努力する

・経営層レベルで握りをつけ、納期調整やトラブル対応の
　幅を広げるようにする
　=＞お互いに組織で対応出来るように

・保守料金には注意　定価が高いと保守も高い
　次年度以降の見積もりを必ず行う
　=＞大事なのは見通しが出来ていること

・コストは月額で試算して比較
　=＞機器費用（リースまたは減価償却）72分割+保守12分割を1ヶ月コストとする

# 機器 -選定　＞対処

・上限値の数字を必ず確認すること

=＞作成するシステムの拡張上限から何が制約になるかを見極める

・検証環境を作ってもらい動作確認をする

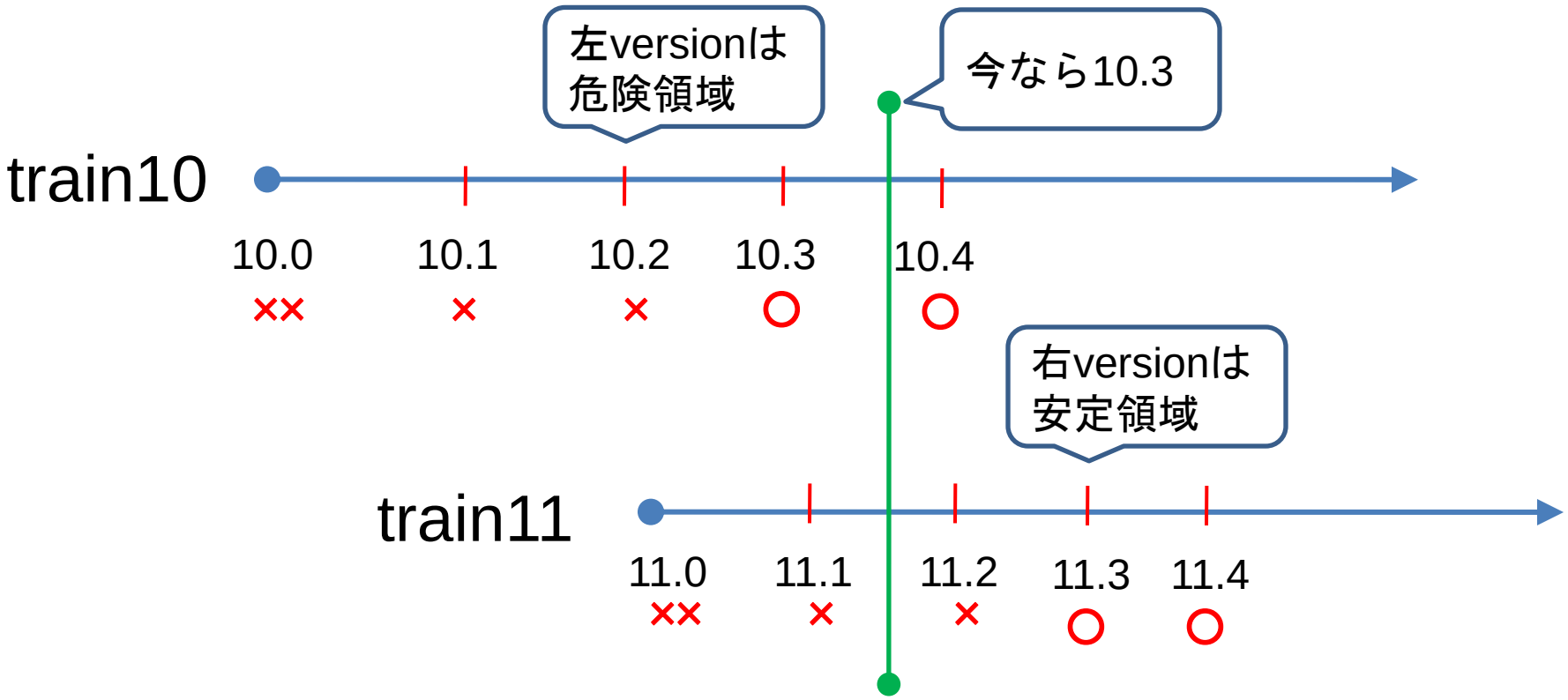
=＞ハードウェアよりファームウェアの機能確認をしっかりすること
　トラブルの大半はファームウェアのバグにより発生する

・トラブルは継続的に発生する　アフターサポートの
　充実したベンダを選ぶ

=＞ネットワーク機器はトラブルや機能拡張の要件は必ず発生し、
　売り切りで終わるものではない
　体力や対応力の高いベンダとパートナー関係を築くこと

# 機器 -ファーム　＞対処

・メジャーリリースの出始めは危険
　ベンダが大丈夫だと言ってもマイナーリリース3くらいまで待つ



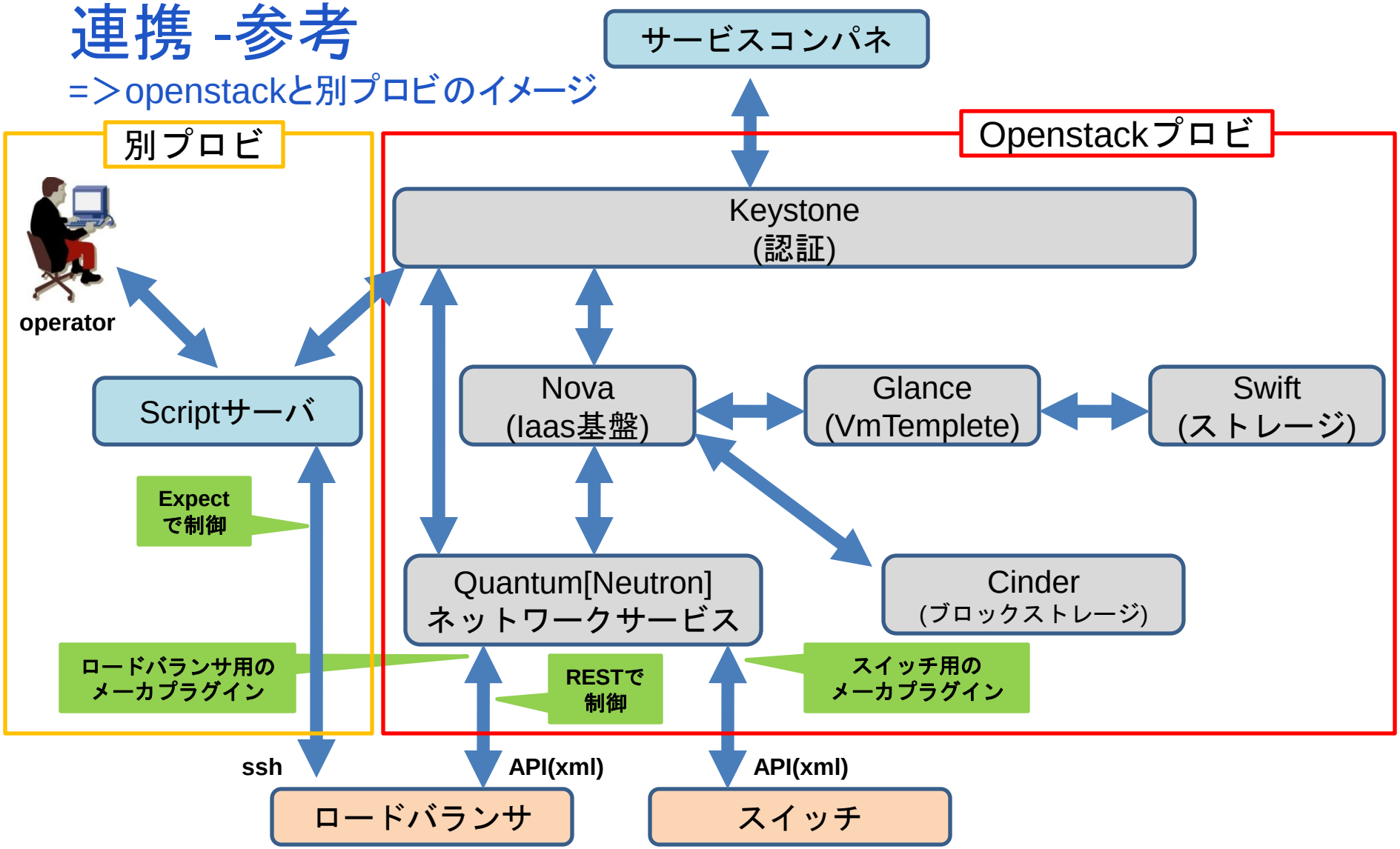
# 連携　＞起きたこと

- ネットワークエンジニアがネットワーク機器の設定を
  行わなずプロビジョニングシステムが自動で行うように
  =＞ API連携やGUIツールからの操作となり、設計とプログラミングに仕事がシフト

- サーバ、プログラマ、ネットワーク、運用等の各チームが
  連携して稼動する機会が増えた
  =＞各チームのスキルの共有などの嬉しい副産物

# 連携 -参考

=＞openstackと別プロビのイメージ



# 連携 >対処

・機器の設定の自動化は出来るところから入る

=> snmp 、excectからtelnet/ssh操作、netconf等色々ありますが
ネットワークエンジニアにとっつき易いのはコンフィグがそのまま見えるexcect

=> phpのモジュールとして利用可能

・Openstackをベースとした構成の場合、Quantum(Neutron)のメーカからリリースされているプラグインを利用する

=>欲しい設定項目が無いとpythonで書かねばならずハードルは高い

=>ネットワーク機器というインフラをに統合してよいのかという議論も

・チームの垣根を超えて連携する

=>ブロビ部分等はネットワークエンジニアだけでは対応が難しい場合が多い
お互いの得意分野を持ち寄りいいシステムを作る

# まとめ =>1

・モバイルサービスの普及により受け皿となる
クラウドサービスの規模は急拡大した
=>競争から高機能、安定性、低コストを求められ、結果インフラの整備が進む

・ピア>>トランジットの流れは加速
=>ビジネス的にクリティカルな通信はトランジットに流さない（ようにしたい）

・クラウドサービスは総力戦で体力勝負
=>システムの数の暴力と競合他社との競争に対応する為、
性能と開発スピードをあげ、少しでもコストと工数を減らす努力を皆が考えることが大切
=>ネットワークエンジニアもコスト意識を持つべし

# まとめ =>2

・機材はシステムの最大拡張時を想定して選ぶ

=>制約事項には注意する

・連携(プロビシステム)は必須となりつつある

=>ネットワークエンジニアという枠に囚われているだけでは生存競争についていけない
それぞれの環境に合った貢献出来ることを考えていく必要がある

=>スピードが重視される為、管理の手間を抑えたり、従来のポリシは柔軟に変えて対応する必要がある

# END OF SLIDE

ご清聴ありがとうございました